ABC HEALTHCARE CASE STUDY

Bhoopesh Sharma

AGENDA

Introduction

Objective

Regression analysis

Outlier Detection

Fraud Detection

Anomalies Detection

Audit Insights and Recommendations

OBJECTIVE

Using descriptive analytics, statistical outlier detection and supervised/unsupervised models to priorities high-risk payments and recommend process & data controls to reduce fraud risk and shorten payment times.

- Detect anomalous/fraudulent payment requests.
- Explain drivers of long payment lead times.
- Deliver actionable sample lists for audit follow-up and controls recommendations.
- Approach (3 pillars):
  - Descriptive & diagnostic analytics (Power BI / Pandas)
  - Statistical outlier detection & sampling (Z-score, IQR, rule-based heuristics)
  - ML & AI (supervised classifier + unsupervised anomaly detection + AI-assisted audit report)

KEY ASSUMPTIONS

- The provided dataset (Payments, Teams, Fraud Cases) is complete, representative, and free from major entry errors.
- All invoices and payments in the dataset are genuine transactions—no synthetic or test data included.
- Invoice IDs, Dates, Amounts, and Team are unique and correctly mapped across all tables.
- Missing or inconsistent values are negligible or imputed using standard statistical methods (e.g., median imputation).
- Payment dates and invoice received dates are accurately recorded in the system (no manual delays or backdating).
- The objective is to identify high-risk invoices, not to make definitive fraud determinations.
- Data refresh cadence for dashboards and models is assumed to be monthly or quarterly.

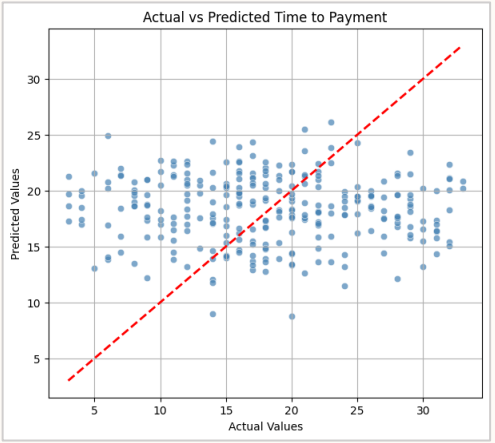
REGRESSION ANALYSIS FOR PAYMENT TIME PREDICTION

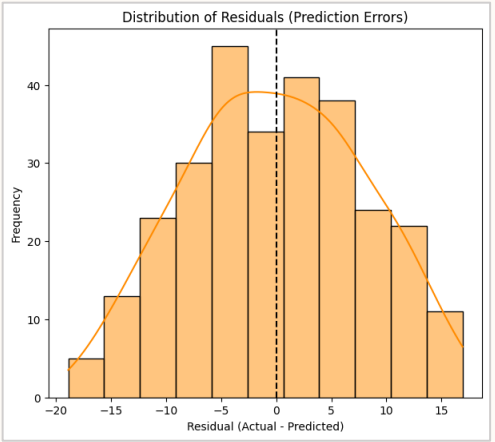
Here are the key results from the Linear Regression model:

- Mean Absolute Error (MAE): **6.68 days**
- Mean Squared Error (MSE): **64.83**
- Root Mean Squared Error (RMSE): **8.05 days**
- R-squared (R2): **- 0.19**

Analysis:

Based on these results, the model does not strongly support the hypothesis that "all invoices take ~same time ±1 day." The Mean Absolute Error of 6.68 days indicates that, on average, our predictions are off by approximately 6.68 days, which is significantly higher than the ±1 day specified in the hypothesis. The negative R-squared value suggests that the model performs worse than a simple horizontal line at the mean of the observed payment times, implying that the current features and linear model are not effectively capturing the underlying patterns in the data.



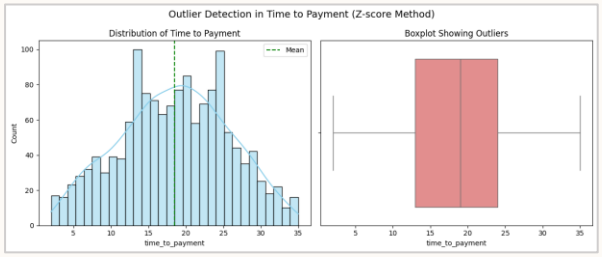


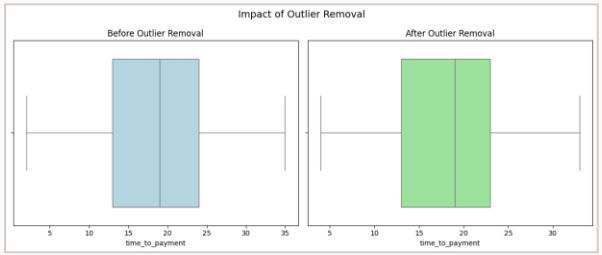
OUTLIER DETECTION USING Z-SCORES FOR AUDIT

Here are the key results:

Identified 33 outliers in 'time_to_payment' using a Z-score threshold of 2.

- Percentage of outliers: **2.31%**
- Summary of time_to_payment:
  - count - **1428.000000**
  - mean - **18.469888**
  - Std - **7.273981**
  - min - **2.000000**
  - max - **35.000000**





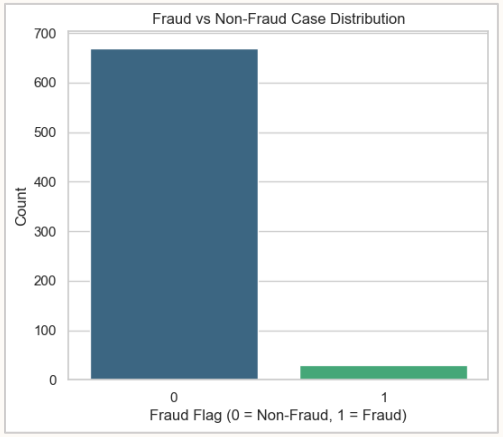
Analysis:

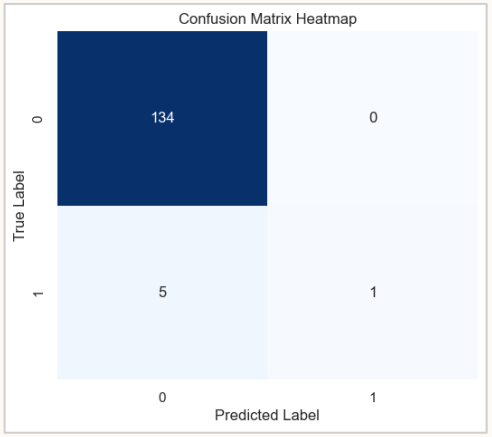
Identified 0 outliers in 'time to payment' using a Z-score threshold of 3 and 33 outliers using 2. This suggests that, based on a normal distribution assumption and these specific threshold, there are no data points that are significantly far from the mean payment time.

It's possible that:

- The time_to_payment data is not normally distributed, making Z-scores less effective for outlier detection.
- The chosen threshold of 3 or 2 is too conservative.
- There are indeed no extreme outliers in the dataset.

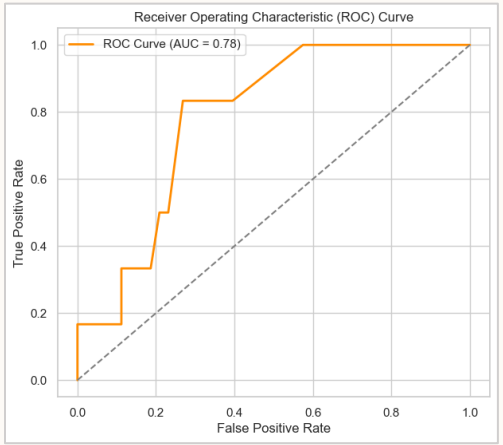
SUPERVISED MODEL FOR FRAUD DETECTION

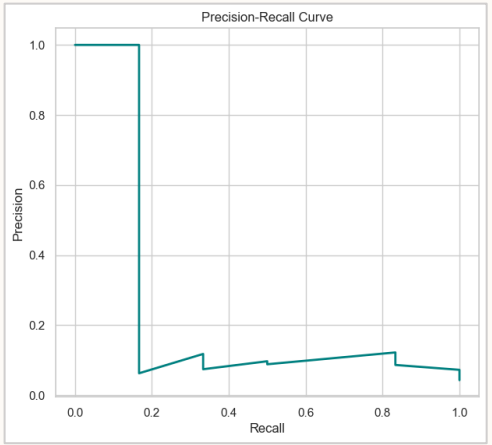


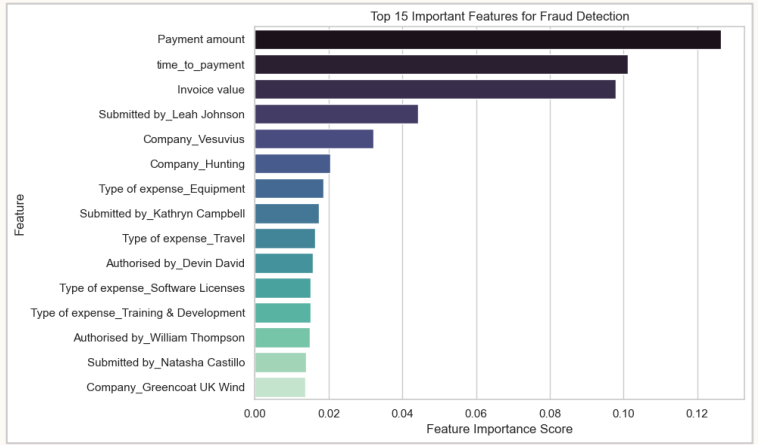


Analysis:

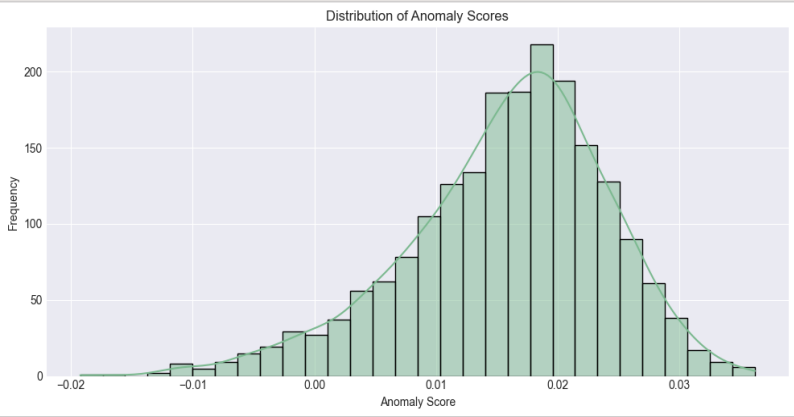
The fraud detection model achieved 96% accuracy and a ROC-AUC score of 0.78, showing a good ability to distinguish between fraud and non-fraud cases. It correctly identified 134 non-fraud and 1 fraud case but missed 5 actual frauds. While its precision for fraud is perfect (1.00), meaning every fraud prediction is correct, its recall is very low (0.17), indicating it detects only 17% of real frauds. The low F1-score (0.29) further reflects this imbalance, suggesting the model struggles to identify all fraudulent transactions due to the highly imbalanced dataset.

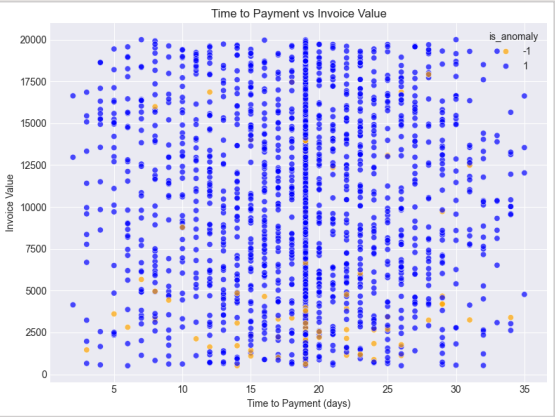


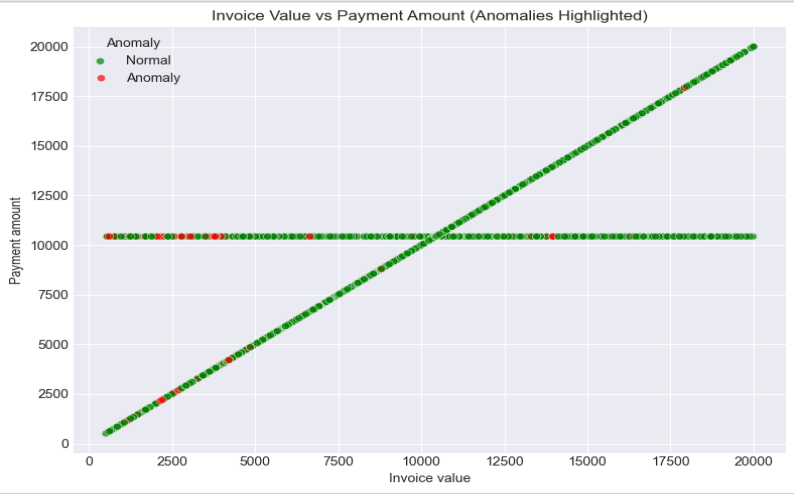


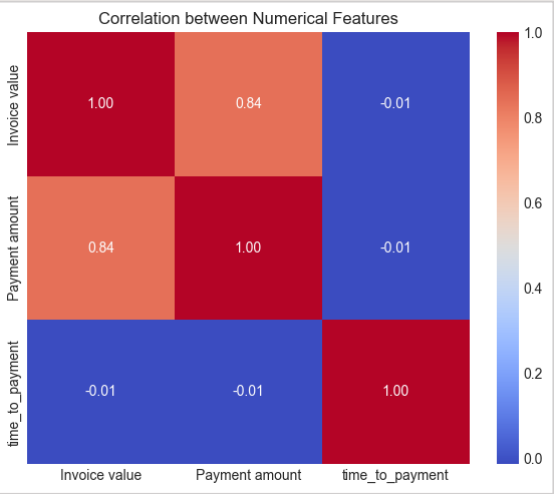


UNSUPERVISED MODEL TO DETECT ANOMALIES









Analysis:

The Isolation Forest model identified 100 instances as anomalies within the payment data, assuming a 5% contamination rate. These anomalies represent data points that deviate significantly from the normal patterns observed in the dataset across various features such as 'Invoice value' and 'time_to_payment'. These identified anomalies warrant further investigation as they could indicate unusual payment activities, potential errors, or even fraudulent transactions that are not easily captured by simple outlier rules.

AUDIT INSIGHTS AND RECOMMENDATIONS

Audit Insight

- Variable Payment Times:
  - High MAE indicates unpredictable payment times.
  - Caused by inefficiencies, varied processes, or missing external factors.
- Limited Z-score Effectiveness:
  - No extreme outliers detected.
  - May be due to data distribution or threshold choice.
  - Suggests need for advanced outlier detection methods.
- Anomalies via Unsupervised Learning:
  - Isolation Forest flagged several statistically unusual transactions.
  - Potential indicators of errors, process deviations, or fraud.
- Incomplete Fraud Detection:
  - Model detected 134 non-fraud and 1 fraud case, missed 5 frauds.
  - Indicates class imbalance issue needing corrective measures.

Recommendations

- Analyze Payment Time Variability *(Identify causes of variation)*
  - Action: Explore key factors affecting payment time (team, expense type, company).
  - Benefit: Highlight process bottlenecks for optimization.
- Improve Outlier Detection *(Capture more unusual payment times)*
  - Action: Test new Z-score thresholds or use IQR for anomaly detection.
  - Benefit: Broader anomaly set for audit review.
- Review Unsupervised Model Anomalies *(Validate accuracy and find hidden irregularities)*
  - Action: Manually verify 100 anomalies flagged by Isolation Forest.
  - Benefit: Early fraud and process issue detection.
- Enhance Supervised Fraud Detection *(Improve fraud prediction accuracy)*
  - Action: Rebuild model with more features and balanced data
  - Benefit: Automate detection, freeing audit resources.

THANK YOU

Bhoopesh Sharma
+91 88261 94431
it.bhoopeshsharma@gmail.com